SAÉ 2.04

Exploitation d'une base de données

Antoine PEREDERII, VIVIEN DUFOUR, CLEMENT LESME
Groupe 9 Première année

SAÉ s2.04 : Exploitation d'une base de données

I. Présentation des jeux de données

A. QUE CONTIENT VOTRE JEU DE DONNEES ?

Le jeu de données disponible ici contient des informations sur les voitures disponibles sur le marché indien. Il a été collecté à partir de CarDekho, une plateforme en ligne de recherche et d'achat de voitures. Le jeu de données est composé de plusieurs fichiers CSV, mais nous allons nous concentrer sur le fichier "car détails v4.csv". Nous avons rajouté deux autres fichiers .csv que nous avons implémentés à la main afin d'avoir d'avantage d'informations sur les marques des véhicules de ce csv.

B. QUELS TYPES DE DONNEES SONT À L'INTERIEUR ?

Ce premier fichier csv "car détails v4.csv" renommé « carDetailsOld.csv » contient des informations détaillées sur 1875 voitures, qui sont décrites dans 20 colonnes :

1. Brand (Marque) - Cette colonne contient le nom de la marque de la voiture.
2. Model (Modèle) - Cette colonne contient le nom du modèle de la voiture.
3. Price (Prix) - Cette colonne contient le prix de vente de la voiture.
4. Year (Année) - Cette colonne contient l'année de fabrication de la voiture.
5. Kilometer (Kilométrage) - Cette colonne contient le kilométrage de la voiture.
6. FuelType (Type de carburant) - Cette colonne contient le type de carburant utilisé par la voiture.
7. Transmission (Transmission) - Cette colonne contient le type de transmission de la voiture.
8. Location (Lieu) - Cette colonne contient lieu de vente da la voiture.
9. Color (Couleur) - Cette colonne contient la couleur de la voiture.
10. Owner (Propriétaire) - Cette colonne contient le nombre de propriétaires précédents de la voiture.
11. SellerType (Type de vendeur) - Cette colonne contient le type de vendeur de la voiture.
12. Engine (Moteur) - Cette colonne contient les détails sur le moteur de la voiture, tels que la cylindrée en cc
13. MaxPower (Puissance maximale) - Cette colonne contient la puissance maximale produite par le moteur de la voiture.
14. MaxTorque (Couple maximal) - Cette colonne contient le couple maximal produit par le moteur de la voiture.
15. Drivetrain (Transmission) - Cette colonne contient le type de transmission de la voiture.
16. Length (Longueur) - Cette colonne contient la longueur de la voiture.
17. Width (Largeur) - Cette colonne contient la largeur de la voiture.
18. Height (Hauteur) - Cette colonne contient la hauteur de la voiture.
19. SeatingCapacity (Capacité d'assise) - Cette colonne contient la capacité d'assises de la voiture.
20. FuelTankCapacity (Capacité du réservoir de carburant) - Cette colonne contient la capacité du réservoir de carburant de la voiture.

Ces colonnes contiennent des données numériques et textuelles, telles que des noms de marques des voitures, leur kilométrage, leur année de fabrication, leurs couleurs, le type de vendeurs, des détails sur les moteurs et la transmission, etc.

Le second fichier CSV "carBrand.csv" contient des informations sur 29 fabricants de voitures différents. Les informations sur chaque fabricant sont stockées dans trois colonnes différentes, à savoir :

1. Name (Nom) - Cette colonne contient le nom du fabricant de la voiture.
2. CreationDate (Date de création) - Cette colonne contient l'année de création du fabricant de la voiture.
3. Headquarter (Siège social) - Cette colonne contient l'emplacement du siège social du fabricant de la voiture.

Ces colonnes contiennent des données textuelles, telles que les noms des fabricants de voitures et leur lieu de sièges sociaux et des données numériques telles que leurs années de création.

Le fichier csv "brandWorkers.csv" contient des informations sur 71 fondateurs et PDG de grandes marques de voitures, décrites dans 4 colonnes :

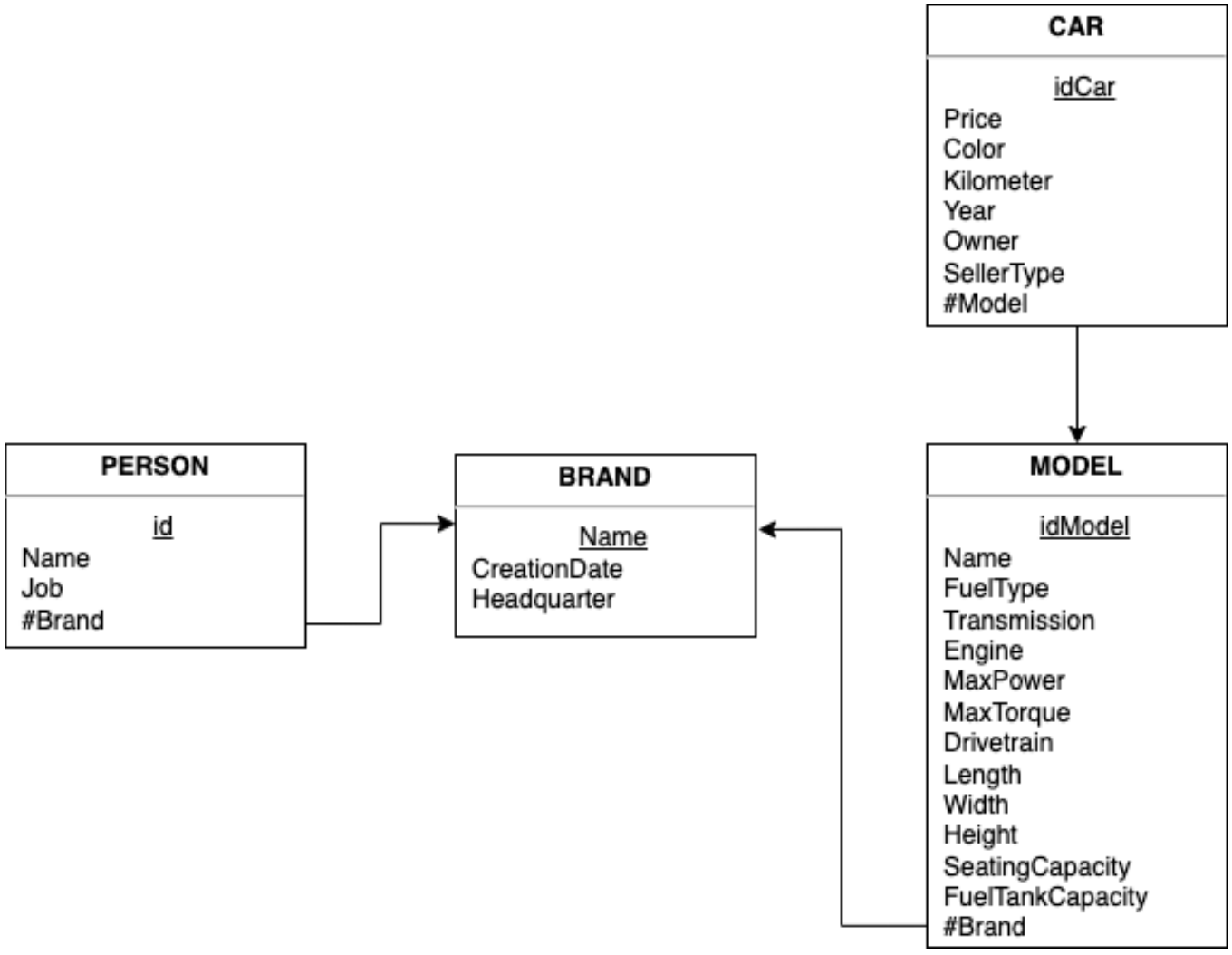
1. Id : identifiant unique pour chaque personne
2. Name : nom de la personne
3. Job : le rôle de la personne dans la société (fondateur ou PDG)
4. Brand : la marque de la voiture associée à la personne.

Le fichier contient donc des données textuelles sur les PDG et Fondateurs des entreprises.

C. QUEL PRÉ-TRAITEMENT AVEZ-VOUS RÉALISÉ (SUPPRESSION DES DOUBLONS, TRANSFORMATION DE VALEURS, ...)

Afin de préparer ce jeu de données pour une analyse ultérieure, plusieurs étapes de prétraitement ont été effectuées, notamment la suppression des doublons, la suppression des valeurs manquantes et la transformation des valeurs de type de carburant en valeurs numériques avec le fichier csvCleaner.ipynb.

II. MLD représentant notre jeu de données :

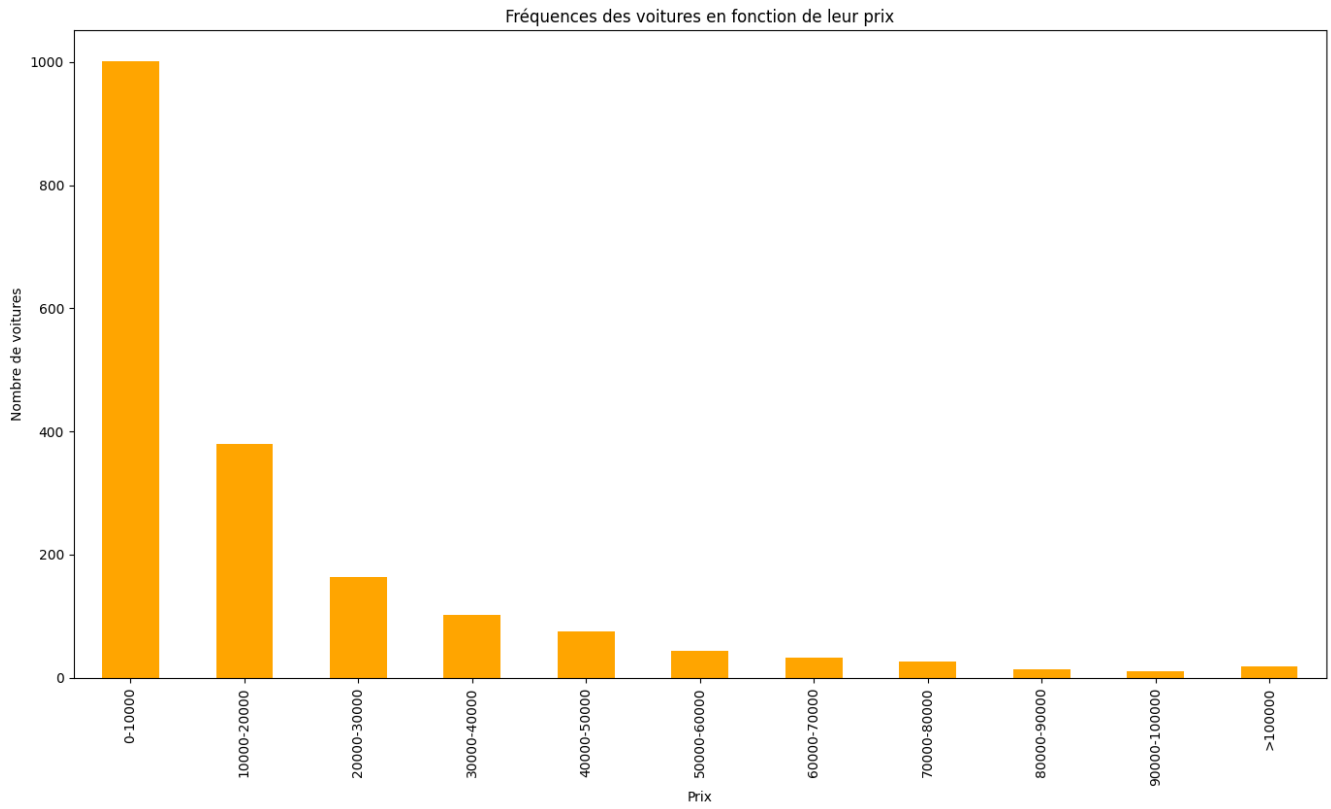


III. Analyse des données

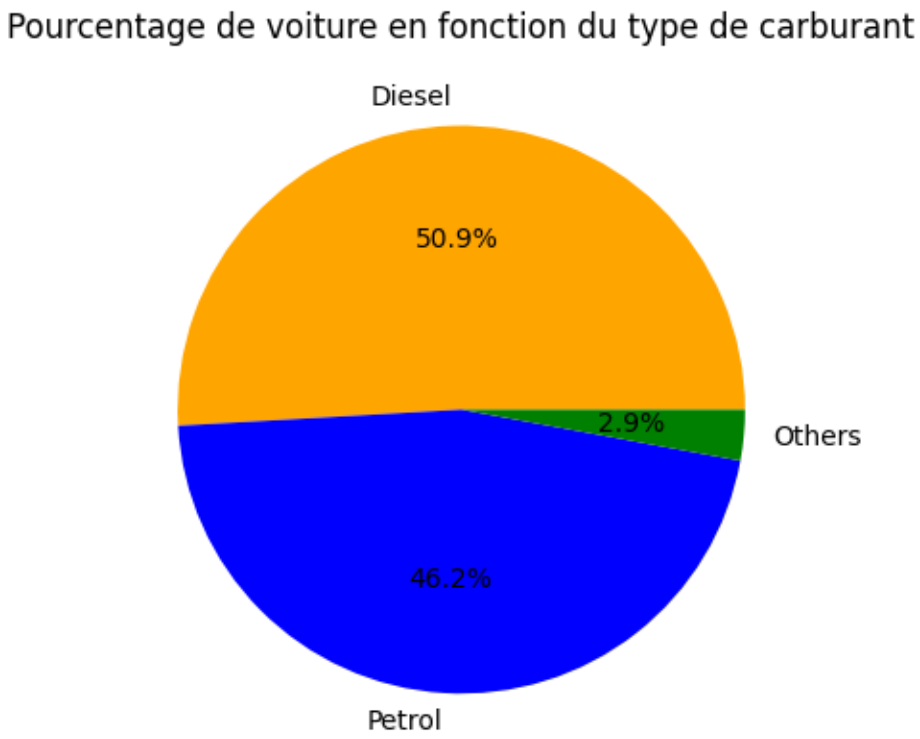
Tout d'abord, nous avons choisi des couleurs de graphique différentes selon notre jugement avant étude sur l'influence des variables au niveau du prix :

- Rouge : Grande
- Orange : Existante
- Bleu : Négligeable

Premièrement, voici un graphique qui permet de répartir le nombre de voitures en fonction de leurs prix afin d'avoir une vue d'ensemble pour la suite :



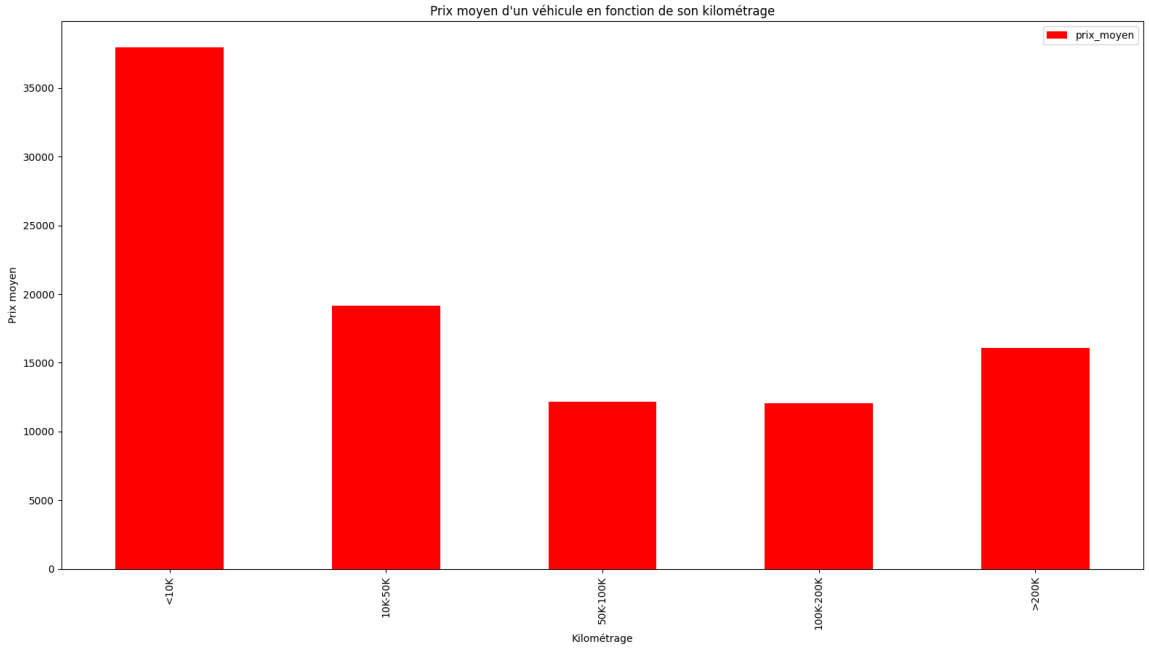
On peut voir que la majorité des voitures de nos données ont un prix inférieur à 10000€ car ce sont souvent des voitures d'occasion ou des voitures peu performantes. Au total, plus de 75% des voitures ont un prix inférieur à 50000€. On peut donc en conclure que moins de 1/4 de ces voitures sont des véhicules haut de gamme de type sport ou lourds comme des pick-up.



La légende est en anglais car notre jeu de données, tout comme notre MLD et nos tables sont dans la langue de Shakespeare.

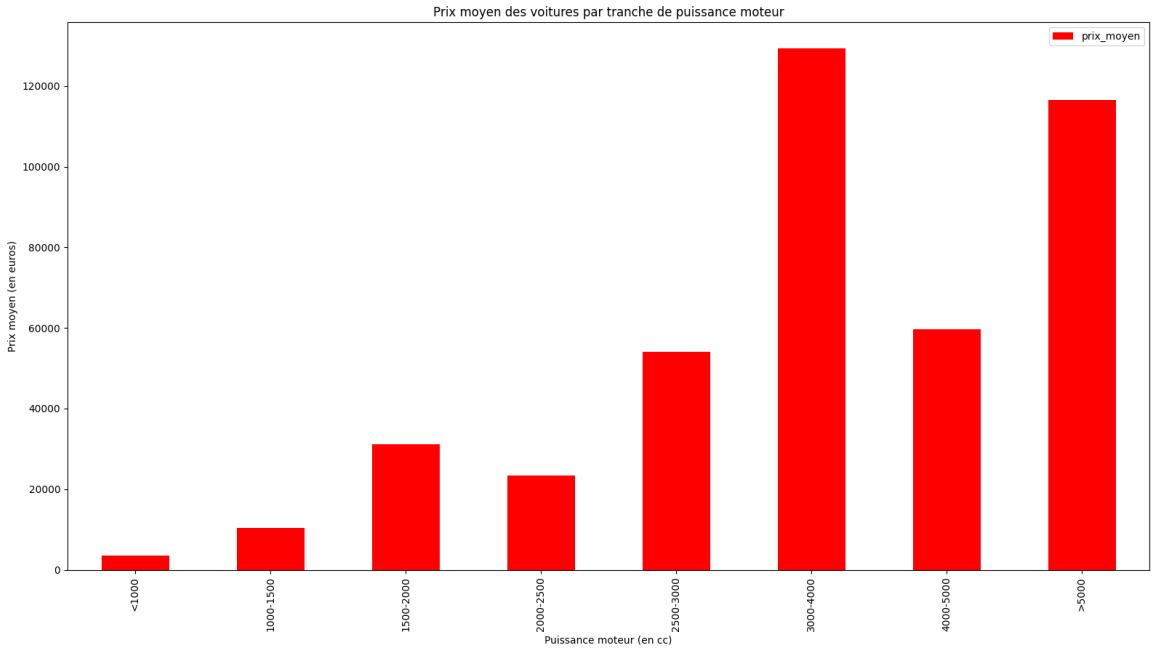
Ici nous avons choisi un diagramme circulaire pour se faire facilement une idée des proportions de chaque type de carburants. On remarque donc que les voitures diesel et essence sont beaucoup plus courantes que les voitures utilisant d'autres types de carburant comme le GPL.

Sur le graphique ci-dessous, on peut voir que le kilométrage d'un véhicule est un facteur clé qui influe sur son prix. Plus le kilométrage est faible, plus le prix est élevé, tandis qu'un kilométrage élevé entraîne une baisse significative du prix. On remarque que le prix entre un véhicule neuf ou avec peu de kilomètres est en moyenne quasiment le double de celui d’un véhicule qui a plus de 10 000 kilomètres au compteur. Cependant, les véhicules ayant un kilométrage très élevé peuvent présenter une valeur légèrement supérieure, car il s’agit généralement de véhicules plus anciens, et donc plus rares sur le marché (véhicules de collection).

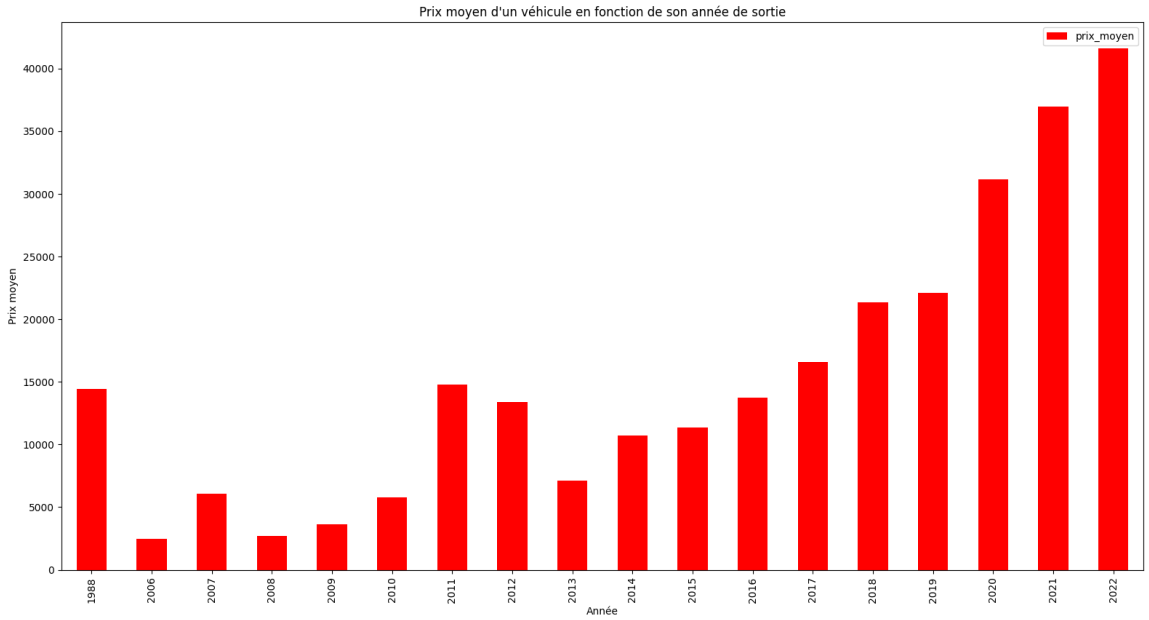


SAÉ s2.04 : Exploitation d'une base de données

Les données du graphique ci-dessous montrent une augmentation des prix plutôt progressive selon la puissance du moteur, puis une ascension fulgurante entre 3000 et 4000 centimètres cubes, certainement due à une ou deux voitures dont le prix est très élevé, brisant l'homogénéité du prix moyen des voitures au niveau de cette tranche.



SAÉ s2.04 : Exploitation d'une base de données



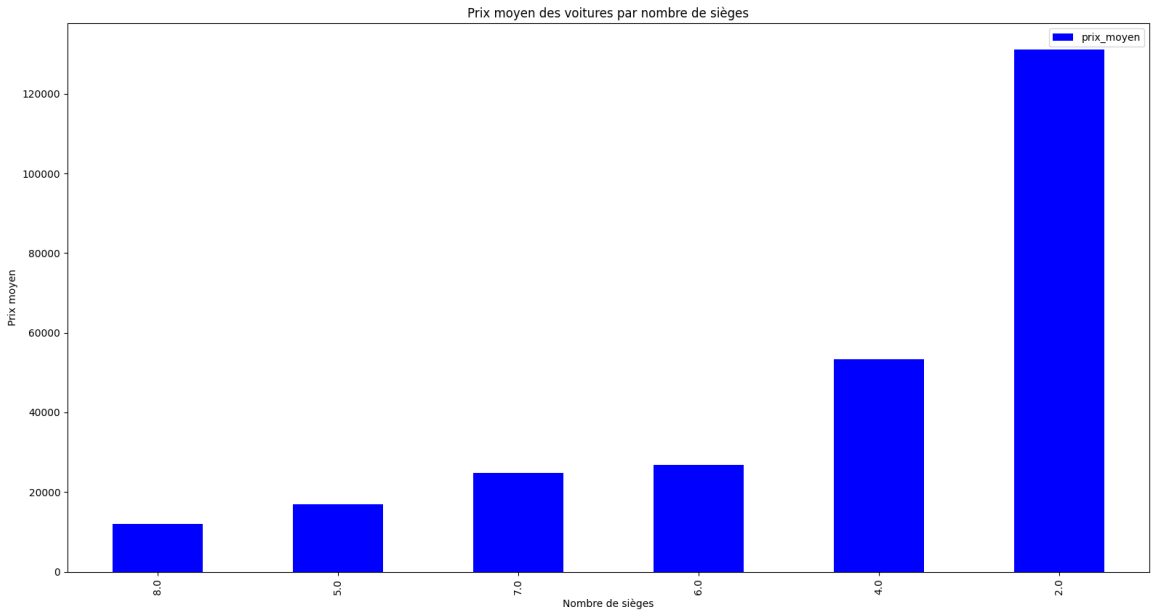
Nous pouvons remarquer, grâce au graphique ci-dessus que la tendance générale est la suivante :

Plus une voiture est récente, plus son prix augmente.

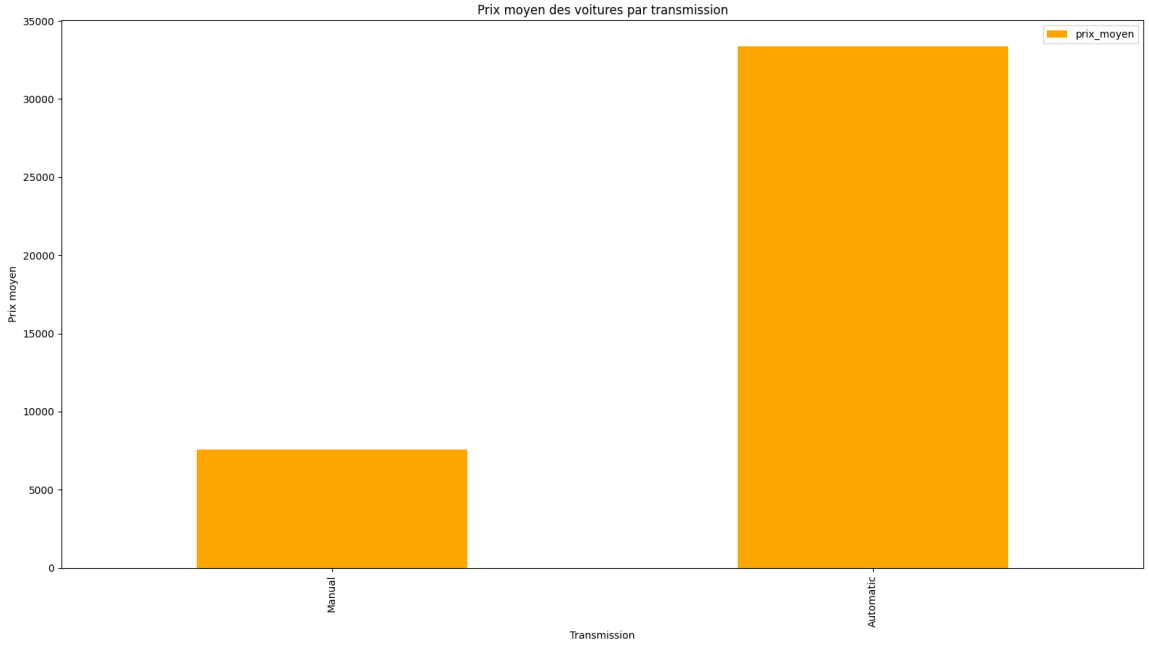
Cependant, cela ne s'applique pas aux véhicules vieux de plus de 2 décennies, car ceux-ci entrent généralement dans la catégorie « Véhicules de collection » ce qui fait nettement augmenter leur prix au vu de leur rareté.

SAÉ s2.04 : Exploitation d'une base de données

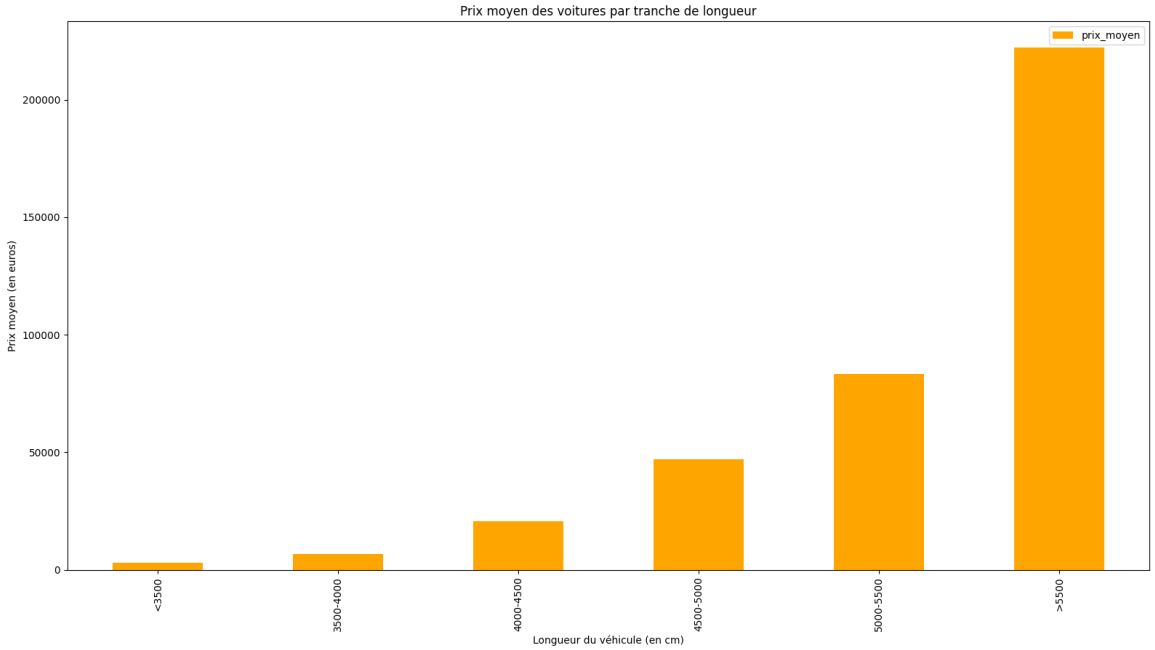
Sur le graphique suivant, on remarque directement qu'une voiture ne possédant que 2 sièges coûte 2 à 4 fois plus chère qu'une voiture possédant plus de places assises, cela est dû au fait que les voitures n'ayant que 2 sièges sont généralement des véhicules de type coupé de sport et qui coûtent généralement plus chers. On peut également voir que c'est assez similaire pour les voitures à 4 sièges qui sont souvent des voitures basiques, mais version sport. Pour les véhicules avec plus de 5 sièges, le prix moyen reste stable car ce sont des voitures de type familiale.



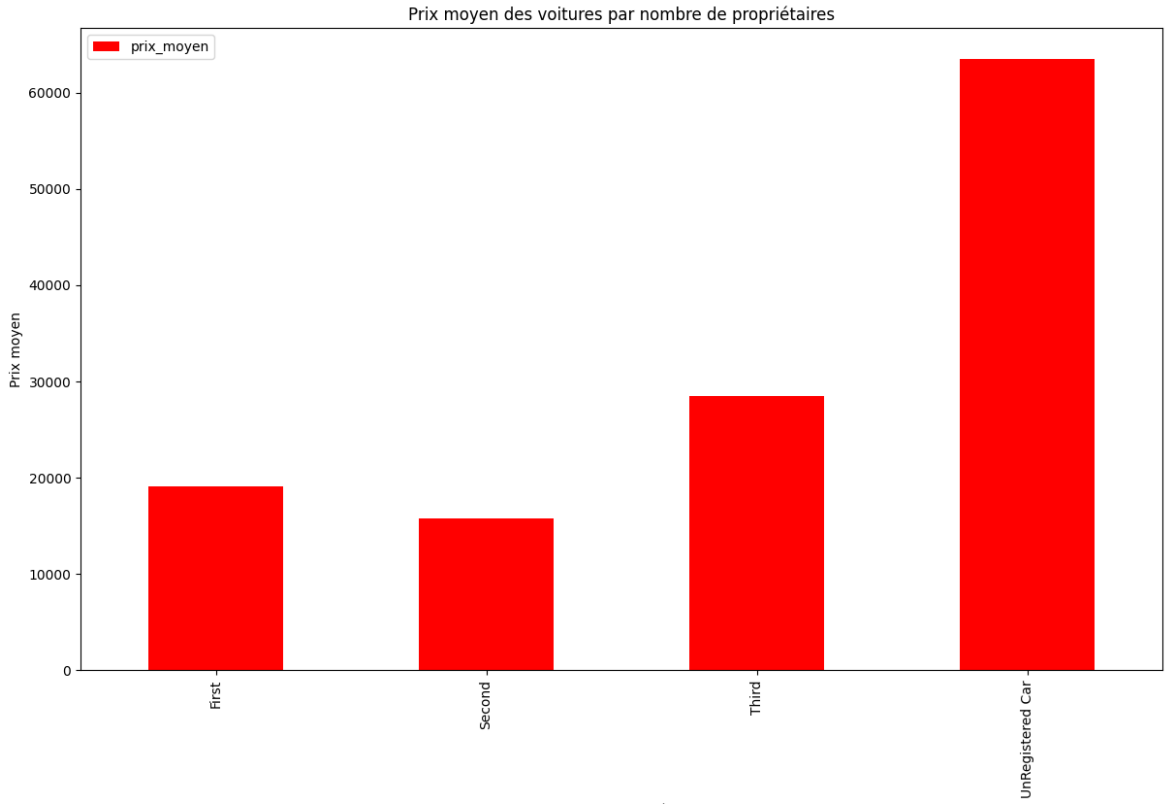
Ici, nous pouvons constater une énorme différence de prix entre les voitures automatiques et manuelles. Les voitures automatiques sont en effet 75 % plus chères en moyenne que les manuelles. Cela est dû au fait que les voitures automatiques sont plus compliquées et donc plus chères à produire. De plus, les boîtes automatiques sont beaucoup plus répandues dans les voitures typées sport que dans les voitures du quotidien.



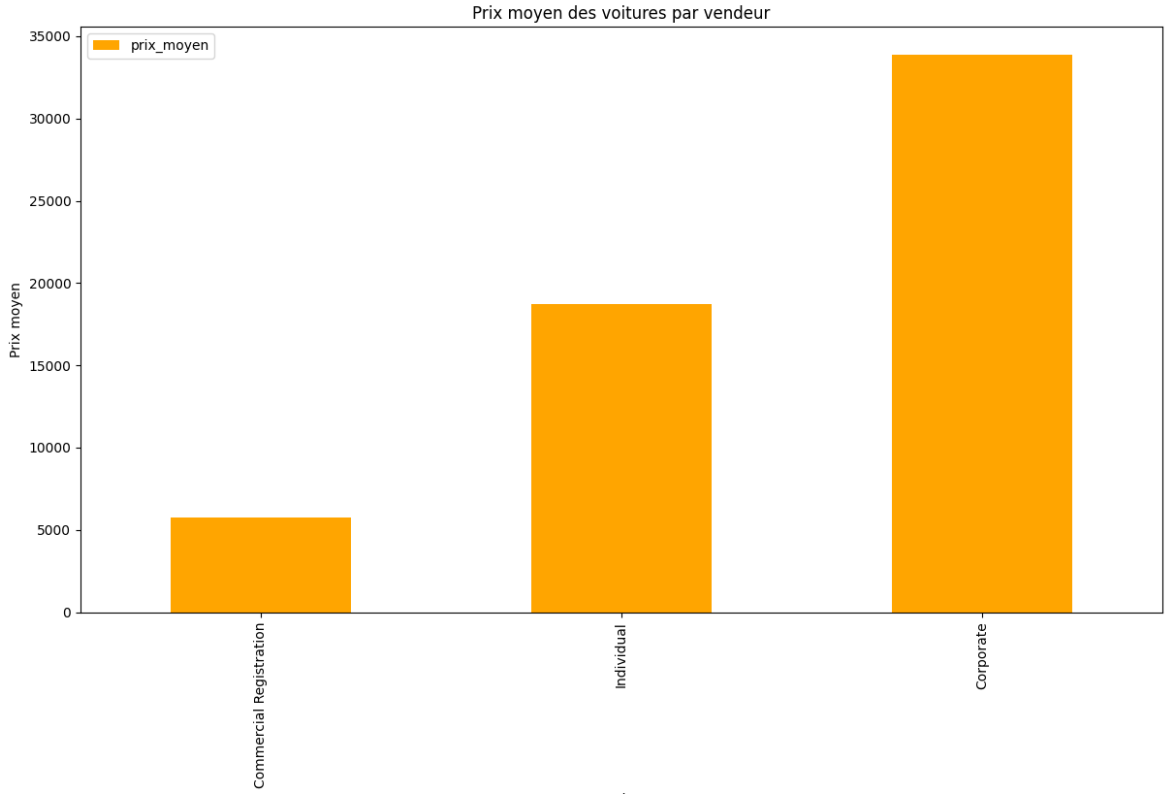
Sur ce graphique, on remarque que la longueur d'un véhicule possède une relative influence sur le prix. Cependant, cette différence devient nettement visible pour des véhicules d'une longueur supérieure à 5,50m certainement due aux voitures plus haut de gamme présentes dans cette tranche. On voit bien que les voitures d'une longueur inférieure à 4m valent nettement moins chères que celles de plus de 5,5m car ce sont généralement des véhicules entrés de gamme, et qui perdent énormément de valeur au cours du temps et de l'usure. De plus, on remarque tout de même qu'une voiture qui fait entre 5m et 5m50 peut coûter 2 à 4 fois plus chère qu'une voiture de plus petite taille.



On remarque ici qu'une voiture neuve coûte le double voire le triple d'une voiture d'occasion. On apprend aussi que même si une voiture a un grand nombre d'anciens propriétaires, son prix baisse, mais bien moins fortement que lorsqu'elle est neuve. On remarque aussi une augmentation du prix des véhicules qui en sont à leur troisième propriétaire, ce qui peut sembler contre intuitif, mais qui, dans les faits, peut de nouveau s'expliquer par la rareté des véhicules plus anciens.



Enfin, nous pouvons constater, à l'aide de ce graphique, que les entreprises vendent les voitures quasiment 2 fois plus chères que les particuliers. Cela reste cohérent car les voitures vendues par des particuliers sont presque toujours d'occasion, là où celles vendues par des entreprises sont soit neuves, soit reconditionnées, ce qui augmente fortement leur prix mais qui apporte une certaine sécurité et sérénité auprès de l'acheteur.



Pour conclure, de nombreux facteurs sont à l'origine du prix des véhicules mais nous pouvons retenir que les produits neufs et plus récents sont toujours préférés. Cependant, de nombreux types de véhicules particuliers modifient grandement les données tels que les véhicules anciens, de collection ou bien les véhicules très haut de gamme qui font considérablement varier les prix des statistiques réalisées.